

SEPTEMBER 2023

AI GENERATION



Render



✕ HACKS OF THE MONTH	02
✕ CYBER NEWS UPDATES	06
✕ JOBS AND INTERNSHIPS	11
✕ QUICK PROJECT	13
✕ ADVANCED PROJECT	19
✕ FACULTY CORNER	26



CAE
IN CYBERSECURITY
COMMUNITY

Digital Shadows: The Rising Tide of Global Cyber Threats

In the past few months, we have witnessed an escalation in cybersecurity threats. Over the past few months, significant incidents have sent ripples through the digital community, reminding us of the ever-present risks. These stories provide a stark revelation: the boundaries between cyber espionage, corporate threats, and state-backed interventions are becoming blurred.

Scraped and Exposed: The Duolingo Debacle

Duolingo's recent incident reveals the dark underbelly of data scraping. When 2.6 million users' data surfaced on a hacking forum, it showcased the evolving strategies of threat actors. This breach wasn't about stealing data but about compiling publicly available information for a nefarious purpose: targeted phishing attacks. Facebook's €265 million penalty, imposed by the Irish Data Protection Commission, reflects the severity of scraped data leaks. This incident tells us that even 'public' data can be weaponized if organized appropriately.

Taiwan in the Crosshairs: Flax Typhoon Strikes

While DuoLingo users were grappling with potential phishing attacks, Taiwan was on the defensive against Flax Typhoon. Microsoft's researchers have traced this hacking group's footprints back to China. Their subtle techniques, relying more on victim systems' inherent tools than malware, show the evolution of stealth in cyber warfare. The implication? Countries must be vigilant against intrusions that might not look like intrusions initially.

Master of Stealth: Flax Typhoon's Evasion

The Flax Typhoon group's insidious methods continued to evolve. Their use of RDP connectivity and legitimate VPN bridges reveals a methodical approach to maintaining long-term access. This isn't a one-time attack; it's a long-game strategy. They fortify their position with each strike, demonstrating an ominously unclear objective.



Digital Shadows: The Rising Tide of Global Cyber Threats

Stars, Stripes, and Satellites: US Space Industry Under Siege

The US is not immune to this wave of cyber threats. The intelligence community's warning of espionage attempts against the American space industry paints a grave picture. The potential threats range from acquiring technological blueprints to jamming satellite communications. The message is clear: As the space race escalates, so will espionage and cyber threats associated with it.

DEF CON: A Mirror to the Cyber World

Amidst these threats, DEF CON serves as a reminder of the incredible talent present in the cybersecurity domain. It's not just about games, fun, or kilts; it reflects the complex cyber landscape. As a gathering point for the world's brightest security minds, DEF CON symbolizes the need for collaboration, innovation, and diligence against emerging threats.

Connecting the Dots

In a few months, we've seen the convergence of different cyber threats. From data scraping for phishing to potential state-backed cyber espionage against strategic sectors, the lines dividing these categories are fading.

There's a collective lesson here: Awareness and Preparedness. Organizations, countries, and individuals must be proactive, anticipate vulnerabilities, and continually fortify defenses. Today's cyber threats might be multi-dimensional, but with a united front, they are not insurmountable.



Digital Warfare on a New Frontier

Over the past three months, a different kind of warfare has surfaced. It's not characterized by ground troops, tanks, or airstrikes but by lines of code, encryption keys, and data breaches. This is cyber warfare, a modern battleground where nation-states, hacktivists, and cybercriminals all stake their claim. Recent hacking incidents involving Russia highlight the deepening complexities of this digital battlefield.

The Battle for the Rails

The hacking collective "Nebula" recently targeted Tehtrans, a Russian corporation tied to the country's railway system, which is implicated in facilitating the movement of military goods. By taking control of the company's website and encrypting its data, Nebula sent a clear message regarding its involvement in Russia's military operations. Such an aggressive maneuver is emblematic of the evolving cyber tactics being deployed by hacktivist groups.

Nebula's action is notable for its historical prowess in targeting Russian governmental systems. However, they are not alone. Another hacking collective, NB65, reportedly linked to the renowned "Anonymous" group, aimed at the Russian space agency, Roscosmos. Their actions disrupted communications and introduced malicious software previously crafted by the Russian government.

Adding to the cyber onslaught, the Russian National Republican Army targeted IT firms associated with Russian military and intelligence services. Their objective? To make these systems more susceptible to future cyber-attacks.

Misinformation and the NATO Summit

On another front, two Russian hacker factions attempted to disrupt NATO's Vilnius summit, a significant diplomatic gathering. By peddling fake NATO press releases through counterfeit websites and social media, they spread disinformation suggesting dramatic shifts in NATO's policy and stance, including doubled defense budgets and the redeployment of Ukrainian troops.



Digital Warfare on a New Frontier



These actions were eerily reminiscent of previous Russian influence campaigns, specifically the "Doppelganger" and "Secondary Infektion" operations. While links between these campaigns and sanctioned Russian entities are yet to be established, the tactics and outcomes are strikingly similar.

However, NATO's response is a beacon for how global organizations should handle such incidents. By swiftly debunking claims and partnering with media outlets to disseminate the truth, NATO exemplified the importance of a proactive stance against cyber threats.

Connecting the Dots

Both stories converge on a singular theme: the sharpening edge of cyber warfare. Once scattered and disorganized, hacktivist collectives now present formidable challenges to nation-states and multinational corporations. The stakes have never been higher, as misinformation campaigns can influence public perception, potentially shifting the balance of power and strategy on the geopolitical stage.

Furthermore, the frequency and intensity of these attacks underscore the dire need for fortified cyber defense systems, collaboration among nations and corporations, and proactive strategies to debunk misinformation swiftly.

In a world where warfare is not confined to physical borders, nations and organizations must remain vigilant against digital threats. The past three months offer a stark reminder: everyone is on the frontline in the digital age.



Exploiting the Windows Filtering Platform: A Deep Dive into Advanced Security Bypass Techniques

The Windows Filtering Platform (WFP) is a set of APIs and system services that provide a platform for creating network filtering applications. Over the years, security researchers have probed its depths, uncovering unique ways to exploit it. Here, we delve into the platform's intricacies, focusing on attackers' advanced techniques and how the platform's security can be bypassed.

The Core of WFP: A Brief Overview

WFP facilitates access to network traffic processing at various TCP/IP protocol stack layers. Critical to its operation is a database known as the policy database, a store for policy rules. Filtering decisions made during network traffic processing are governed by these rules, which administrators and applications can configure.

Duplicating Tokens via WFP

The Technique

The ability to duplicate process tokens is a crucial avenue for privilege escalation attacks. Typically, by duplicating a SYSTEM token, an attacker can elevate a process to run with higher privileges. In the context of WFP:

1. Using **NtQueryInformationProcess**, an attacker can retrieve the handle table of another process, which provides a list of tokens held by that process.
2. By interfacing with the kernel, this technique duplicates tokens without calling **DuplicateHandle** from user mode, thus evading detection by most EDR solutions.
3. Key to this is using a Device IO request, calling **WfpAleProcessTokenReference** which then duplicates a SYSTEM token and stores it in a hash table.

Significance and Detection

This technique offers a unique method to duplicate tokens of several services, such as LSM, Winmgmt, and Schedule. Defending against such methods requires vigilant monitoring of suspicious actions, especially when processes running with lower permissions attempt to duplicate handles to a SYSTEM token.

Exploiting the Windows Filtering Platform: A Deep Dive into Advanced Security Bypass Techniques

Triggering IPsec Connection for Exploitation

The Technique

Internet Protocol Security (IPsec) is a suite of protocols ensuring secure and private network communication. When leveraged maliciously:

1. An attacker can force a service (like the Print Spooler service) to initiate a connection that matches an IPsec policy. This results in the insertion of a SYSTEM token into a table.
2. RPC methods, such as **RpcOpenPrinter**, are then exploited to force a service connection to a socket.

Significance and Detection

This approach is stealthy. Configuring an IPsec policy is a standard action often performed by network administrators. Most EDR solutions monitoring network activity might overlook connections to localhost, making this attack especially dangerous.

Manipulating User Services for Token Acquisition

The Technique

RPC servers provide a ripe ground for token manipulation. The process is as follows:

1. Identify RPC servers running as logged-on users.
2. Target services like OneSyncSvc that load libraries such as SyncController.dll.
3. Exploit RPC methods to force service connections and capture tokens.

Significance and Detection

Exploiting services like OneSyncSvc and SyncController.dll is groundbreaking because offensive tools have not previously targeted them. As a result, security solutions may overlook such invocations.

Exploiting the Windows Filtering Platform: A Deep Dive into Advanced Security Bypass Techniques

Further Intricacies and Potential Attacks

In-depth research into the tcpip.sys driver reveals additional functionalities exposed through device IO requests. Some of the potentially exploitable functionalities relate to:

- Hash table operations and managed data.
- Process Explicit Credentials stored in the tcpip driver.

Exploiting process-explicit credentials can provide attackers with obscure and undocumented avenues for data retrieval. However, this area remains less understood and requires further exploration.

Conclusion

The Windows Filtering Platform is a complex beast, ripe for exploitation by those with the knowledge and intent. Recent discoveries in its vulnerabilities have shown the potential for lateral movement, privilege escalation, and stealthy attacks that bypass conventional security mechanisms. As these techniques evolve, so too must our understanding and defenses. One crucial takeaway is the need for persistent, in-depth security research that continuously challenges the boundaries of existing systems. This ensures that platforms like WFP remain robust against the ever-evolving threat landscape.

Note: All system administrators and security professionals must stay updated on these findings and promptly apply patches or mitigation techniques as recommended by software vendors.



HiatusRAT: An Escalating Cyber Threat with Geo-Political Undertones

A clandestine hacking operation, Hiatus, has been quietly accumulating victims and evolving tactics over the past year. Leveraging a sophisticated cocktail of malware components, this campaign primarily targets business-class VPN routers and can steal data, surveil networks, and build a covert proxy network. New reports suggest a strategic shift in its targeting preferences with geopolitical implications.

Technical Overview

The Hiatus campaign relies on three main components: a malicious bash script, a malware named HiatusRAT, and the legitimate 'tcpdump' utility. Once gaining unauthorized access to targeted DrayTek Vigor routers, commonly used by small to medium-sized organizations, the malware starts listening on port 8816 and collects extensive information, ranging from system data and networking data to file system and process data. It also sends a heartbeat POST to its command and control (C2) server every 8 hours, effectively allowing the attackers to keep tabs on compromised routers.

Evolving Techniques and Scale

HiatusRAT has shown its resilience by recompiling malware samples to target different architectures, including Arm, Intel 80386, x86-64, MIPS, MIPS64, and i386. Despite previous reporting that may have exposed its operations, the group has only made minor modifications to their payload servers and continued to function nearly unabated.

Geopolitical Shifts

The latest campaign has witnessed a significant change in target preferences. Whereas initial reports indicated primary targeting in Europe, North America, and South America, recent findings suggest a focus on Taiwanese organizations and even a U.S. military procurement system. These shifts coincide with strategic interests identified in the 2023 ODNI threat assessment, linking them to the People's Republic of China.

Implications

These tactical shifts are not just about expanding the target list. They have significant consequences. For instance, the data transfer connections have shifted from primarily Latin American and European entities to Taiwanese organizations and entities linked with the U.S. Department of Defense. This suggests potential intelligence gathering aligned with geopolitical interests.

HiatusRAT: An Escalating Cyber Threat with Geo-Political Undertones

Conclusion

Due to the evolving nature and the potential geopolitical implications of this threat, both businesses and consumers must remain vigilant. Comprehensive Secure Access Service Edge (SASE) solutions and the latest cryptographic protocols are strongly advised for bolstering network security. Moreover, regular monitoring, rebooting, and installing security updates are recommended for self-managed routers.

Given the complexity and audacity displayed by the HiatusRAT campaign and its geopolitical undertones, this constitutes a threat that requires the collective attention of the cybersecurity community and policymakers alike.

The revelation of HiatusRAT's recent activities, particularly its strategic shift towards targeting Taiwanese organizations and a U.S. military procurement system, intensifies suspicion between the U.S. and China. This comes when cyber-espionage and cyber-warfare have become pivotal points of contention in international relations. The alignment of HiatusRAT's activities with the strategic interests identified in the 2023 ODNI threat assessment further implicates the People's Republic of China, whether or not direct state sponsorship can be proven.

Such revelations could strain the tenuous U.S.-China relations, further complicating diplomatic engagements and trade negotiations. The cybersecurity breaches, especially those aimed at military systems, not only pose a direct threat to national security but also reinforce the narrative of China as an aggressive actor in the cyber realm. This may push the U.S. to adopt a tougher stance, potentially catalyzing the implementation of sanctions, joint cybersecurity initiatives with allies, or even establishing a clearer cyber doctrine. Concurrently, China would be expected to deny such allegations, framing them as geopolitical maneuvers to contain China's rise. These dynamics, born from the digital shadows, might lead to tangible real-world implications in the diplomatic arena between the two superpowers.



Student Intern, IT Operations and Enterprise Security

Tucson Electric Power, a UNS Energy Corporation subsidiary, has served communities since the 1890s. They are driven by a vision of providing clean, green, and reliable electric and gas services to nearly 700,000 customers in Arizona. The company values its employees, offering a competitive compensation package, opportunities for professional growth, a collaborative work environment, and a commitment to the betterment of the communities they serve.

Job Description:

The Student Intern will work in IT Operations and Enterprise Security, assisting the management team in various capacities. Responsibilities include budgeting, tracking and analyzing expenses, data analysis, monthly financial reporting, and other ad-hoc reporting needs.

Ideal Candidate:

Actively enrolled in college, pursuing a degree in Finance, Accounting, Business Administration, or a related field, with a graduation date of May 2025 or later.

Proficient in Microsoft Office, especially Excel.

Familiar with Business Intelligence tools such as Oracle BI and Microsoft Power BI.

Strong interpersonal and communication skills.

Ability to handle confidential information, analyze data, and work both independently and in a team.

Committed to working up to 20 hours during regular school weeks and up to 40 hours during breaks, as needed.

Considerations for Students:

This position is an excellent opportunity for students seeking practical experience in a professional setting, especially those interested in IT operations, enterprise security, and financial analysis. Beyond gaining industry-specific knowledge, interns will also be part of a company that values its workforce and actively contributes to the community. The role offers a competitive pay rate and the chance to work for a reputable utility company in Arizona. Students who are proactive, keen to learn, and can balance their academic commitments with work should consider applying.



Cybersecurity Analyst Spring 2024 Intern - Remote

Southwest Airlines is renowned for its commitment to providing employees with a stable work environment that fosters learning and personal growth. They encourage creativity and innovation, especially when it comes to enhancing the efficacy of the airline. They prioritize providing employees with the same respect, concern, and caring attitude they expect employees to extend to Southwest Customers.

INTERNSHIP DESCRIPTION:

As a Cybersecurity Analyst Intern, you will join the Incident Response Team to tackle real-world cyber challenges. This position allows you to gain hands-on experience in identifying, mitigating, and recovering from cyber threats. You'll have the chance to collaborate with professionals to assess threats, strategize responses, and bolster defense mechanisms.

KEY RESPONSIBILITIES:

Providing technical support to diagnose and counteract software/hardware issues.
Engaging in service desk functions, ticket inquiries, and troubleshooting.
Assisting in system software and equipment setup, configuration, and installation.
Programming and troubleshooting computer software and hardware.
Assisting in evaluating and recommending new software applications and equipment implementation.

IDEAL CANDIDATE:

Recent graduate or college student majoring in Cyber Security, MIS, Computer Science, or related studies.
Demonstrates proficiency in Microsoft Office Suite and has a basic understanding of information security.
Experience or exposure to tools like Proofpoint PhishAlarm and ServiceNow Governance, Risk, and Compliance (GRC) is a plus.
Strong communication skills, both verbal and written.
Ability to manage multiple tasks, work under tight deadlines, and demonstrate effective problem-solving capabilities.
Should have a sense of humor, be team-oriented, and be a quick learner.

WHAT THE INTERNSHIP OFFERS:

12-week internship running from January 30th to April 19th, 2024.
Fully remote work, with potential opportunities for virtual team interaction and a possible visit to the headquarters.
Hourly pay of \$30, with a one-time \$500 stipend on the first paycheck.
Unlimited travel privileges for the intern.
Required equipment (e.g., laptop) provided for the internship.
Part-time commitment between 24-40 hours a week.

Considerations for Students:

This internship is a stellar opportunity for students looking to immerse themselves in the world of cybersecurity. Working with Southwest Airlines, a respected name in the airline industry, interns get to understand the nuances of cybersecurity in a real-world context. The internship is designed to empower students with the skills required in crisis management, threat assessment, and collaboration. Those keen on diving deep into the realm of cybersecurity, learning from seasoned professionals, and making a tangible difference should seriously consider this position. The added perks of travel privileges and a competitive stipend make it all the more attractive.

Southwest
Careers



JOBS & INTERNSHIPS

THE UNIVERSITY
OF ARIZONA 12

Discover SearXNG: Your Local Search Engine

Search engines act as our guiding star in the vast expanse of the internet. But did you know you can have your search engine in your local environment? Enter SearXNG, a privacy-respecting, hackable metasearch engine. By the end of this article, you'll have your instance of SearXNG running. So, gear up, and let's embark on this thrilling adventure!

What is SearXNG?

SearXNG is a fork of Searx, a respected metasearch engine. It brings together results from various search engines and presents them to you. But what makes SearXNG shine is its emphasis on privacy, the ability to customize, and the user-centric improvements over its predecessor.



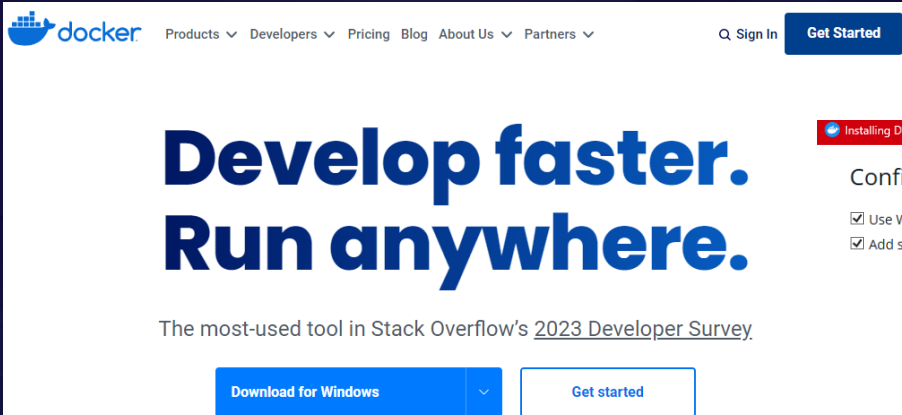
To set this up, we need to ensure that Docker is installed. I am using Windows 11, and to install Docker, I need to run the command `wsl --update` to ensure that your Windows Subsystem for Linux is current and enabled.

```
Command Prompt - wsl --up  X + v
Microsoft Windows [Version 10.0.22621.2215]
(c) Microsoft Corporation. All rights reserved.

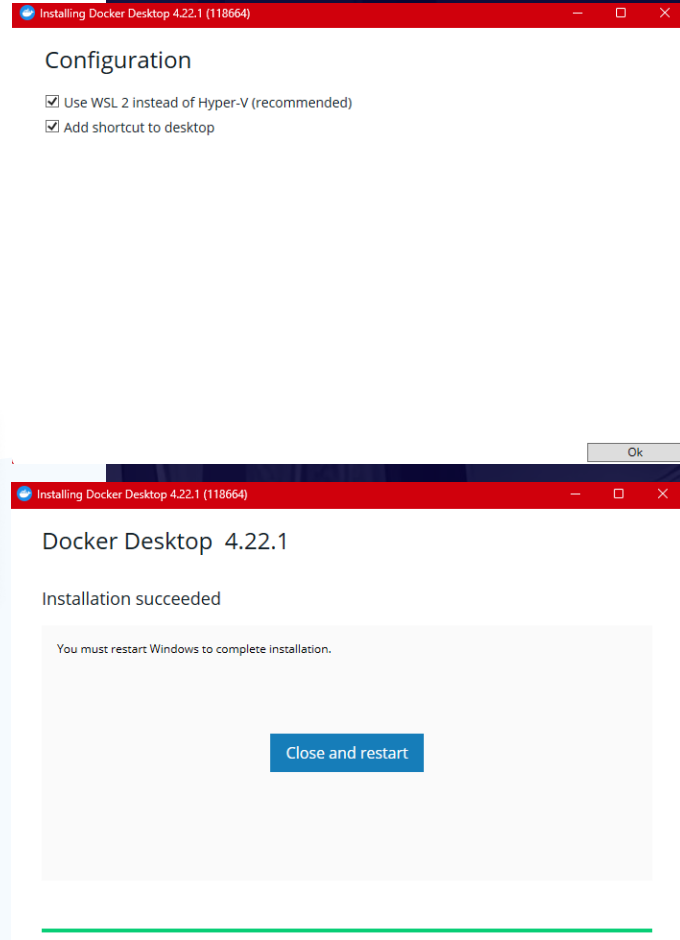
C:\Users\micha>wsl --update
Installing: Windows Subsystem for Linux
[==                               4.0% ]
```


Discover SearXNG: Your Local Search Engine

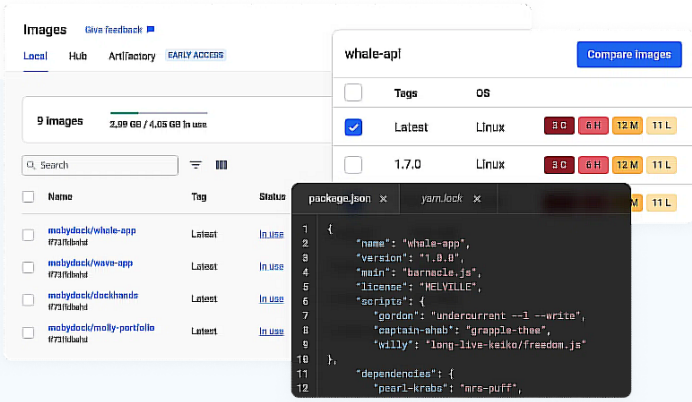
Once that is complete, we need to navigate to docker.com and download a copy of Docker Desktop for Windows.



The screenshot shows the Docker website homepage. At the top, there is a navigation bar with links for Products, Developers, Pricing, Blog, About Us, and Partners, along with a search icon and 'Sign In' and 'Get Started' buttons. The main headline reads 'Develop faster. Run anywhere.' Below this, it states 'The most-used tool in Stack Overflow's 2023 Developer Survey'. There are two buttons: 'Download for Windows' and 'Get started'.



This block contains two screenshots of the Docker Desktop installation process. The first window is titled 'Installing Docker Desktop 4.22.1 (118664)' and shows a 'Configuration' section with two checked options: 'Use WSL 2 instead of Hyper-V (recommended)' and 'Add shortcut to desktop'. The second window is titled 'Installing Docker Desktop 4.22.1 (118664)' and shows 'Docker Desktop 4.22.1' with the message 'Installation succeeded'. Below the message, it says 'You must restart Windows to complete installation.' and there is a 'Close and restart' button.



The screenshot shows the Docker Hub interface for the 'whale-api' image. It displays a table of images with columns for Name, Tag, Status, and OS. A 'package.json' file is overlaid on the image, showing the image name and version. The package.json content is as follows:

```
1 {
2   "name": "whale-api",
3   "version": "1.0.0",
4   "main": "barneacle.js",
5   "license": "MIT",
6   "scripts": {
7     "gordon": "undercurrent --l --write",
8     "captain-shab": "grapple-three",
9     "willy": "long-live-keiko/freedom.js"
10  },
11   "dependencies": {
12     "pearl-krabs": "mrs-puff",
```

What is Docker

Trusted by developers Chosen by Fortune 100 companies

Docker provides a suite of development tools, services, trusted content, and automations, used individually or together, to accelerate the delivery of secure applications.

Once we finish the installation and restart Windows, we can run the Docker program. We are then directed to create/log into a Docker account, and we are able to see Docker Compose.

Discover SearXNG: Your Local Search Engine

At this point, we are ready to get our copy of SearXNG from the docker repository. First, we will create a folder to hold our image. I named mine “my-instance” and moved it into this new folder. I do this with the following command after opening the Windows command prompt.

```
Mkdir my-instance
```

```
Cd my-instance
```

We can now run the docker command to pull the latest image from the repository by running the following command:

```
Docker pull searxng/searxng
```

We will allow this to pull the image, which may take a moment to transfer to your machine.

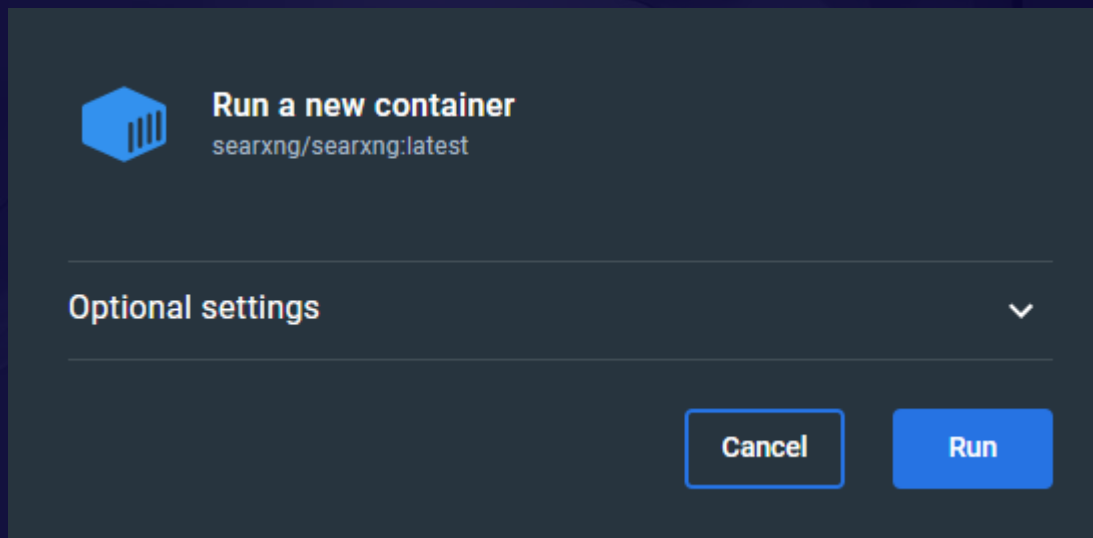
```
C:\Users\micha\my-instance>docker pull searxng/searxng
Using default tag: latest
latest: Pulling from searxng/searxng
7264a8db6415: Pull complete
f5f28345052c: Pull complete
4f4fb700ef54: Pull complete
e0f708a4919d: Pull complete
2b2437bbdf89: Downloading [=====>] 12.43MB/63.81MB
4b5f8157a007: Download complete
d85a000df733: Download complete
332069420cb8: Download complete
```

When this is complete, we can open the docker desktop, navigate to the images section, and find our image.

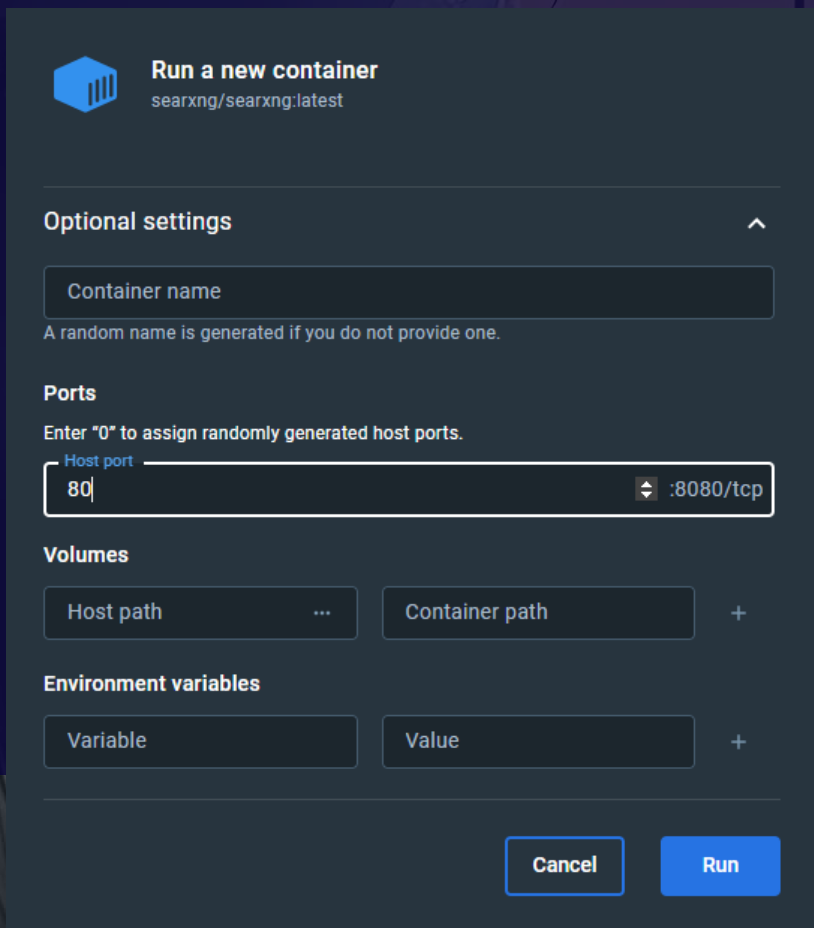
Name	Tag	Status	Created	Size	Actions
searxng/searxng	latest	In use	6 hours ago	213.96 MB	

Discover SearXNG: Your Local Search Engine

Now, we click on the image and open it up. We are shown a screen asking us to run a new container like the one below.



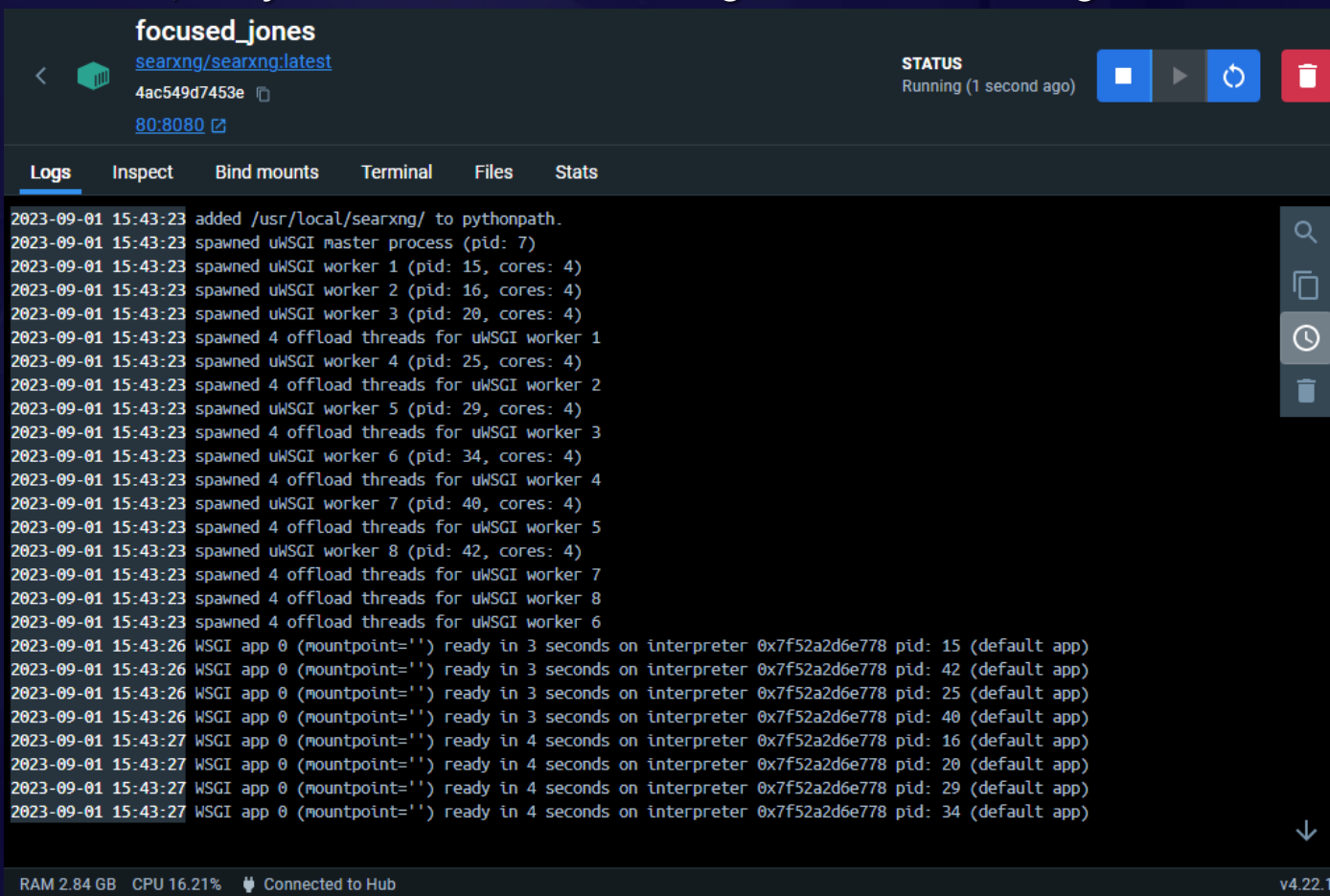
Before we do this, however, we need to tell docker we want to open up a port so we can view the website. We do this by clicking on Optional settings.



We then assign 80 to the Host port. This allows us to view the site using `http://localhost` on our web browser of choice.

Discover SearXNG: Your Local Search Engine

Click on run and allow the image to start up. This usually only takes a few moments, but you should see something similar to the image below.

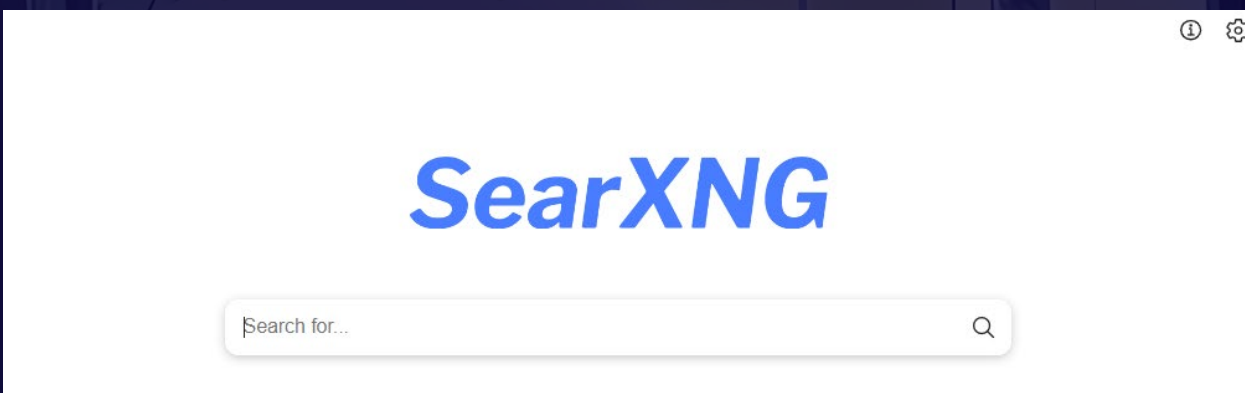


The screenshot shows a Docker container named 'focused_jones' running 'searxng/searxng:latest'. The status is 'Running (1 second ago)'. The logs show the following sequence of events:

```
2023-09-01 15:43:23 added /usr/local/searxng/ to pythonpath.
2023-09-01 15:43:23 spawned uWSGI master process (pid: 7)
2023-09-01 15:43:23 spawned uWSGI worker 1 (pid: 15, cores: 4)
2023-09-01 15:43:23 spawned uWSGI worker 2 (pid: 16, cores: 4)
2023-09-01 15:43:23 spawned uWSGI worker 3 (pid: 20, cores: 4)
2023-09-01 15:43:23 spawned 4 offload threads for uWSGI worker 1
2023-09-01 15:43:23 spawned uWSGI worker 4 (pid: 25, cores: 4)
2023-09-01 15:43:23 spawned 4 offload threads for uWSGI worker 2
2023-09-01 15:43:23 spawned uWSGI worker 5 (pid: 29, cores: 4)
2023-09-01 15:43:23 spawned 4 offload threads for uWSGI worker 3
2023-09-01 15:43:23 spawned uWSGI worker 6 (pid: 34, cores: 4)
2023-09-01 15:43:23 spawned 4 offload threads for uWSGI worker 4
2023-09-01 15:43:23 spawned uWSGI worker 7 (pid: 40, cores: 4)
2023-09-01 15:43:23 spawned 4 offload threads for uWSGI worker 5
2023-09-01 15:43:23 spawned uWSGI worker 8 (pid: 42, cores: 4)
2023-09-01 15:43:23 spawned 4 offload threads for uWSGI worker 7
2023-09-01 15:43:23 spawned 4 offload threads for uWSGI worker 8
2023-09-01 15:43:23 spawned 4 offload threads for uWSGI worker 6
2023-09-01 15:43:26 WSGI app 0 (mountpoint='') ready in 3 seconds on interpreter 0x7f52a2d6e778 pid: 15 (default app)
2023-09-01 15:43:26 WSGI app 0 (mountpoint='') ready in 3 seconds on interpreter 0x7f52a2d6e778 pid: 42 (default app)
2023-09-01 15:43:26 WSGI app 0 (mountpoint='') ready in 3 seconds on interpreter 0x7f52a2d6e778 pid: 25 (default app)
2023-09-01 15:43:26 WSGI app 0 (mountpoint='') ready in 3 seconds on interpreter 0x7f52a2d6e778 pid: 40 (default app)
2023-09-01 15:43:27 WSGI app 0 (mountpoint='') ready in 4 seconds on interpreter 0x7f52a2d6e778 pid: 16 (default app)
2023-09-01 15:43:27 WSGI app 0 (mountpoint='') ready in 4 seconds on interpreter 0x7f52a2d6e778 pid: 20 (default app)
2023-09-01 15:43:27 WSGI app 0 (mountpoint='') ready in 4 seconds on interpreter 0x7f52a2d6e778 pid: 29 (default app)
2023-09-01 15:43:27 WSGI app 0 (mountpoint='') ready in 4 seconds on interpreter 0x7f52a2d6e778 pid: 34 (default app)
```

At the bottom, system resources are shown: RAM 2.84 GB, CPU 16.21%, and the container is connected to a Hub. The version is v4.22.1.

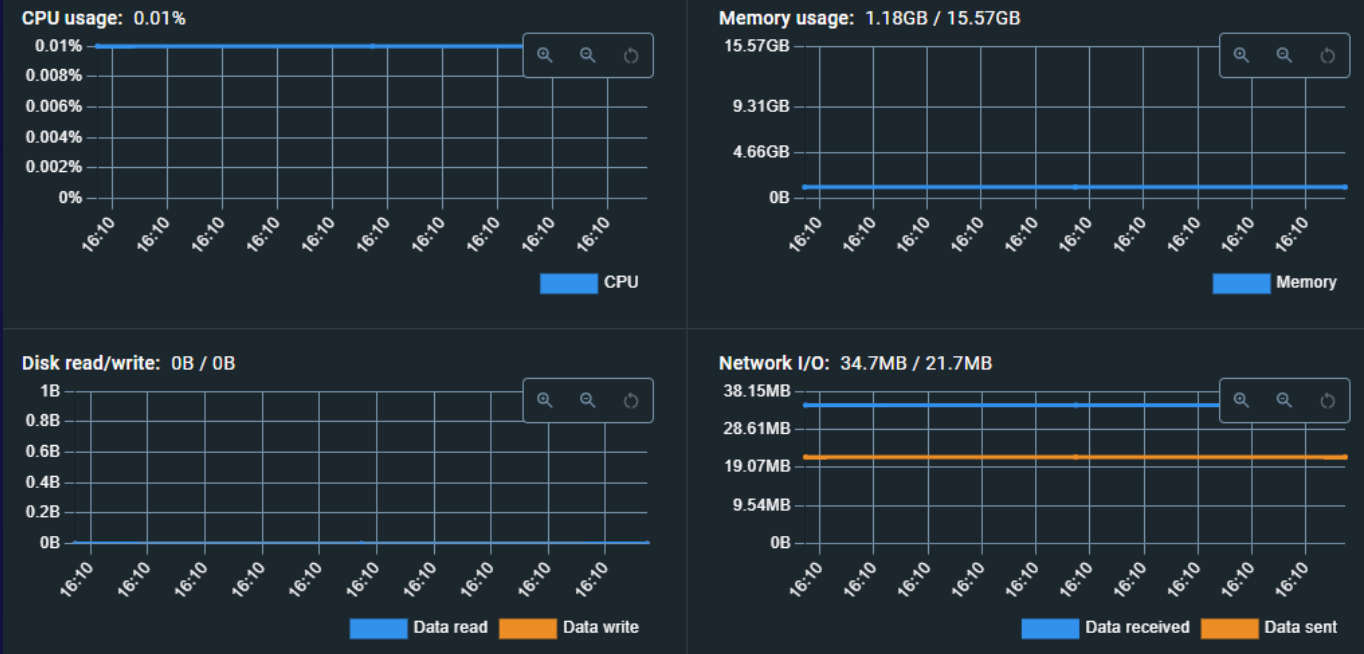
Next, open up your web browser and observe your very own local web browser that allows you to control your data.



Now you can search the internet while keeping your information locally and configure your search engine. I changed the logo on my start page as an example but there are many other options available to you as well.

Discover SearXNG: Your Local Search Engine

The search engine running within Docker is lightweight and should not cause many issues, in my testing I did not encounter any performance issues, but depending on your machine you may notice some slowness.



If you experience lag or performance issues, you can reset to the default base image by removing your private container and initiating it once more. While this means losing your configurations, the convenience of quickly setting up your search instance, especially if you prioritize privacy or data concerns, makes this approach valuable.

The image shows three screenshots of search results for 'University of Arizona' on the SearXNG search engine. The top screenshot is a full-page view showing search filters (General, Images, Videos, News, Map, Music, IT, Science, Files, Social Media), language (English [en]), and search results from various sources like www.arizona.edu and en.wikipedia.org. The middle screenshot is a vertical list of video results from YouTube, including 'University of Arizona Campus Tour', 'Make Yourself at the University of Arizona', and 'Tour of The University of Arizona for International Students'. The right screenshot shows a vertical list of map results from OpenStreetMap, including 'University of Arizona', 'University of Arizona Global Center', 'University of Arizona Police', and 'University of Arizona - Chandler'.

Introducing PrivateGPT: A Local Language Model for Cybersecurity Enthusiasts

In the era of cloud computing, data privacy has taken a backstage for many applications. This project introduces "PrivateGPT," a project that brings the power of advanced language models to your local machine, ensuring 100% data privacy. Tailored for cybersecurity students in this project, this tool can be a stepping stone in understanding the crossroads of AI and cybersecurity.

What is PrivateGPT?

PrivateGPT is an innovative project that allows users to interact with a language model locally without any internet connection. By leveraging the capabilities of Large Language Models (LLMs), it enables users to ask questions to their documents and receive answers, all while ensuring that no data ever leaves the local environment.

Why Should Cybersecurity Students Care?

- **Data Privacy:** Data breaches and unauthorized data access are among the biggest threats in the modern age. By running AI models locally, students can appreciate the importance of data privacy and learn to build solutions that prioritize it.
- **Hands-on Experience with AI:** AI and cybersecurity are interlinked domains. By setting up and using PrivateGPT, students get hands-on experience with cutting-edge AI tools.
- **Custom Datasets:** PrivateGPT supports various document types, from text files to PowerPoints, allowing students to ingest and question their datasets.
- **Understand Underlying Mechanisms:** By delving into the code and architecture of PrivateGPT, students can understand how embeddings, vector stores, and LLMs work together to provide answers.

How Does PrivateGPT Work?

The magic behind PrivateGPT is a combination of several tools:

- **Document Parsing & Embedding Creation:** The ingest.py script leverages tools to parse documents and create embeddings. These embeddings are stored in a local vector database.
- **Question Understanding & Answering:** The privateGPT.py script uses a local LLM to understand questions and generate answers. It extracts the context for answers from the local vector store, ensuring relevant responses.

Introducing PrivateGPT: A Local Language Model for Cybersecurity Enthusiasts

System Requirements:

- **Python Version:** PrivateGPT requires Python 3.10 or later.
- **C++ Compiler:** Some installations may need a C++ compiler, especially during pip installation.

PrivateGPT will stand at the intersection of AI and cybersecurity for this project. For students looking to get their hands dirty with real-world tools while understanding the significance of data privacy, PrivateGPT is a project worth exploring. So, roll up your sleeves, dive into the world of local language models with PrivateGPT, and get a glimpse of the future of private AI!

To start our adventure, we need to set up our environment. We would need to ensure that Python and Visual Studio Code are installed with C++ for a Windows machine.

Python

Visual Studio Code

Visual Studio Build Tools

Now that we have those, we need to download and extract the GitHub project locally.

GitHub Project

Now, we can run the command **pip3 install -r requirements.txt** to install the project requirements onto the system.

```
PS C:\Users\micha\Downloads\privateGPT-main\privateGPT-main> pip3 install -r requirements.txt
Collecting langchain==0.0.274 (from -r requirements.txt (line 1))
  Obtaining dependency information for langchain==0.0.274 from https://files.pythonhosted.org/packages/ce/e1/59fc4dbe3a72be422c62ca96e670d590a0f7b6c795fdd0c3d757736a1a19/langchain-0.0.274-py3-none-any.whl.metadata
  Downloading langchain-0.0.274-py3-none-any.whl.metadata (14 kB)
Collecting gpt4all==1.0.8 (from -r requirements.txt (line 2))
  Obtaining dependency information for gpt4all==1.0.8 from https://files.pythonhosted.org/packages/55/04/a03ca8551103ad1852ealbb7257b63f0be0d8e5d909fa8fa484efaac0729/gpt4all-1.0.8-py3-none-win_amd64.whl.metadata
  Downloading gpt4all-1.0.8-py3-none-win_amd64.whl.metadata (894 bytes)
Collecting chromadb==0.4.7 (from -r requirements.txt (line 3))
  Obtaining dependency information for chromadb==0.4.7 from https://files.pythonhosted.org/packages/bd/47/17b44d8c372d32ec8cf1901801e163e960859b9e610cc623a3350afe8924/chromadb-0.4.7-py3-none-any.whl.metadata
  Downloading chromadb-0.4.7-py3-none-any.whl.metadata (6.9 kB)
Collecting llama-cpp-python==0.1.81 (from -r requirements.txt (line 4))
  Downloading llama_cpp_python-0.1.81.tar.gz (1.8 MB)
  ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 1.8/1.8 MB 2.0 MB/s eta 0:00:00
Installing build dependencies ... /
```

Introducing PrivateGPT: A Local Language Model for Cybersecurity Enthusiasts

With the project requirements loaded, we need to install the language model for our project. There are a few to choose from, but we will download the following into the same directory for this project.

GPT4ALL LLM

Once we have moved the LLM into our directory, we will update the local environment file by editing `example.env` with a text editor.

Follow the GitHub page to modify this file but the one I am using for my examples is as follows:

```
PERSIST_DIRECTORY=db
MODEL_TYPE=GPT4All
MODEL_PATH=ggml-gpt4all-j-v1.3-groovy.bin
EMBEDDINGS_MODEL_NAME=all-MiniLM-L6-v2
MODEL_N_CTX=1000
MODEL_N_BATCH=8TARGET_SOURCE_CHUNKS=4
```

After completing this, we want to move this into the project environment by running the following command.

```
cp example.env .env
```

Next, we want to find some documents to ingest into our project, and I have chosen a selection of PDF documents to work against. The project GitHub claims that Microsoft Office document files with the DOCX and PPTX extensions work however, I have not had success with these. Text and PDF files, however, have been able to work, so feel free to play around with them. The project includes a text document of the 2022 State of the Union address.

Introducing PrivateGPT: A Local Language Model for Cybersecurity Enthusiasts

I decided to include a few of my lecture PDF slides in the source document section to include these in my analysis.

Name	Date modified	Type	Size
Today			
state_of_the_union	9/6/2023 9:22 AM	Text Document	39 KB
Last month			
CYBV 330 Fall Syllabus 2023	8/7/2023 1:24 PM	Adobe Acrobat D...	309 KB
Earlier this year			
2023 CYBV 326 Lecture 13	4/18/2023 10:10 AM	Adobe Acrobat D...	3,889 KB
2023 CYBV 326 Lecture 9	3/21/2023 11:09 AM	Adobe Acrobat D...	4,583 KB
2023 CYBV 326 Lecture 7	2/28/2023 11:34 AM	Adobe Acrobat D...	6,374 KB
2023 CYBV 326 Lecture 5	2/14/2023 10:21 AM	Adobe Acrobat D...	4,690 KB
2023 CYBV 326 Lecture 3	1/31/2023 9:18 AM	Adobe Acrobat D...	4,585 KB
2023 CYBV 326 Lecture 1	1/16/2023 3:47 PM	Adobe Acrobat D...	6,361 KB
A long time ago			
2022 CYBV 326 Lecture 13	11/15/2022 9:14 AM	Adobe Acrobat D...	4,096 KB
2022 CYBV 326 Lecture 11	11/1/2022 9:22 AM	Adobe Acrobat D...	3,759 KB
2022 CYBV 326 Lecture 9	10/18/2022 9:06 AM	Adobe Acrobat D...	5,357 KB
2022 CYBV 326 Lecture 7	10/4/2022 8:36 AM	Adobe Acrobat D...	5,479 KB
2022 CYBV 326 Lecture 5	9/20/2022 1:59 PM	Adobe Acrobat D...	4,235 KB
2022 CYBV 326 Lecture 3	9/6/2022 8:35 AM	Adobe Acrobat D...	4,883 KB
2022 CYBV 326 Lecture 2	8/30/2022 8:45 AM	Adobe Acrobat D...	571 KB
2022 CYBV 326 Lecture 1	8/23/2022 9:49 AM	Adobe Acrobat D...	5,071 KB
CYBV 330 Introduction to Industrial C...	12/14/2020 2:23 PM	Adobe Acrobat D...	203 KB
CYBV 330 Introduction to Industrial C...	12/14/2020 2:13 PM	Adobe Acrobat D...	1,636 KB
CYBV 330 Introduction to Industrial C...	12/14/2020 2:11 PM	Adobe Acrobat D...	1,418 KB
CYBV 330 Introduction to Industrial C...	12/14/2020 2:01 PM	Adobe Acrobat D...	1,400 KB
CYBV 330 Introduction to Industrial C...	12/14/2020 1:08 PM	Adobe Acrobat D...	1,648 KB
CYBV 330 Introduction to Industrial C...	12/14/2020 1:01 PM	Adobe Acrobat D...	1,766 KB
CYBV 330 Introduction to Industrial C...	12/14/2020 12:59 PM	Adobe Acrobat D...	2,406 KB
CYBV 330 Introduction to Industrial C...	12/14/2020 12:11 PM	Adobe Acrobat D...	3,492 KB

Why Do Token Limits Exist

Efficiency and Performance: Handling vast tokens in one go can be resource-intensive and slow down the model's response time. By setting a token limit, developers ensure that the model operates efficiently, providing timely responses.

How Do Token Limits Impact LLM's Memory

Temporary Memory Storage: Imagine conversing with someone with a short-term memory. They can recall only the last few statements you made. Similarly, an LLM has a "memory" defined by its token limit. If the model's token limit is 500, it can "remember" or consider only the last 500 tokens in its interactions. This mechanism helps the model maintain context during a conversation.

Implications for User Interactions

Maintaining Context: Due to the token limit, if a user's query or the conversation history exceeds the model's token capacity, older tokens (or earlier parts of the conversation) get discarded. As a result, if crucial context is lost, the user might need to repeat or rephrase their question to get a relevant answer. For instance, if you had a long chat and asked, "What about it?" the model might not recall the initial topic if it's beyond its token memory.

In essence, token limits are a trade-off between performance and context retention in LLMs. While they ensure efficient operation, they also constrain how much context or conversation history the model can retain at any given time. Keep this in mind when you are pushing your documents to be ingested and what questions you are asking using this project. A single document with 100 pages would only utilize a portion of the ingesting methods used in this project.

This page by itself utilizes 419 tokens of just this text. You can use tools like [Tokenizer](#) to help clarify what portion of your document is covered by the ingester database-building limitations. A helpful rule of thumb is that one token generally corresponds to ~4 text characters for common English text. This translates to roughly $\frac{3}{4}$ of a word (100 tokens \approx 75 words).

Introducing PrivateGPT: A Local Language Model for Cybersecurity Enthusiasts

Now that we have explained a little about tokens, we can understand how to “break” up our data and ensure that our ingesting covers everything we intend to review. The more data we ingest, the larger our database will be. Keep this in mind as you add more documents to be included.

We can now ask our first question by running the following command

```
python .\privateGPT.py
```

```
PS C:\Users\micha\Downloads\privateGPT-main\privateGPT-main> python .\privateGPT.py
Found model file at C:\Users\micha\Downloads\privateGPT-main\privateGPT-main\ggml-gpt4all-j-v1.3-groovy.bin
gptj_model_load: loading model from 'C:\Users\micha\Downloads\privateGPT-main\privateGPT-main\ggml-gpt4all-j-v1.3-groovy.bin' - please wait ...
gptj_model_load: n_vocab = 50400
gptj_model_load: n_ctx = 2048
gptj_model_load: n_embd = 4096
gptj_model_load: n_head = 16
gptj_model_load: n_layer = 28
gptj_model_load: n_rot = 64
gptj_model_load: f16 = 2
gptj_model_load: ggml ctx size = 5401.45 MB
gptj_model_load: kv self size = 896.00 MB
gptj_model_load: ..... done
gptj_model_load: model size = 3609.38 MB / num tensors = 285
```

```
Enter a query: What does CYBV 330 refer to
```

This will load in the chosen language model, connect the database, and allow us to ask questions against our data. It takes over a minute, but we get our results. Now, I can use the power of AI against my data and this program even provides sources for the information.

```
> Question:
```

```
What is a PLC and what does it stand for?
```

```
> Answer (took 117.91 s.):
```

```
A Programmable Logic Controller (PLC) stands for "Programmable Logical Control." It's an industrial automation device that can be programmed to control various machines or processes in the manufacturing, transportation, logistics, healthcare, retail, energy, water treatment, food and beverage, chemical processing, mining, oil refining, power generation, waste management, environmental remediation, and other industries. PLCs are used for a wide range of applications such as controlling conveyor belts, elevators, escalators, door openers, forklifts, robotic arms, welding machines, drilling equipment, water treatment plants, chemical processing units, oil refineries, power generation stations, waste management facilities, environmental remediation sites, and many more.
```

```
> source_documents\CYBV 330 Introduction to Industrial Control System Security - Lecture 1.pdf:
```

```
Difference between IT & OT
```

- PROGRAMMABLE LOGIC CONTROLLER
- The workhorse of the industrial automation space, PLCs

Privacy Data Mappings BY: Professor Jordan A. VanHoy

Allow me to introduce you to a paper written by one of our own professors, Professor VanHoy who wrote the article Privacy Data Mappings. In an era of heightened concern for personal data privacy, organizations worldwide grapple with understanding and managing the vast arrays of data they possess. Despite the introduction of robust regulations such as the European General Data Protection Directive (GDPR) and the California Consumer Privacy Protection Act (CCPA), many entities remain unaware of the intricacies of their data management. This paper delves into the crucial role of data mapping in bridging this gap. Data mapping, the process of identifying and understanding personal data flow within an organization's systems, proves instrumental in complying with intricate regulatory requirements. The paper underscores the benefits of streamlined, efficient data mapping processes by exploring various frameworks and techniques, including automated solutions. Using real-world cases such as Doorstep Dispensaree Ltd's GDPR violation, the study emphasizes the urgent need for organizations to adopt comprehensive data mapping to ensure data privacy and regulatory compliance.

Data Mapping

Throughout the global economy there exists systemic issues regarding the privacy of collected personal information. Recent regulations have attempted to curb this issue such as the European General Data Protection Directive (GDPR) and the California Consumer Privacy Protection Act (CCPA). However, it is still all too common to find that the average organization has little to no idea what data they have, why they have specific data, or how long the data has resided with the company, among many questions to be answered. A method that may be employed by organizations to begin rectifying this is to conduct data mapping. Data mapping is the process of identifying and understanding of how personal data is used in an organization's information systems.

To highlight this issue, Doorstep Dispensaree Ltd received penalties for violating the GDPR in 2019 for failing to properly secure and protect the personal information of their data subjects. During the Medicines and Healthcare products Regulatory Agency's search, permitted by a warrant, they identified 47 crates, two disposal bags and one cardboard box full of documents containing personal data consisting of approximately 500,000 documents relating to an unknown number of data subjects and therefore notified the Information Commissioners Office of a possible GDPR violation (Boger, 2020, p.1). The information included names, addresses, dates of birth, National Health Service numbers, medical information and prescriptions belonging to an unknown number of people (DataGuidance, 2022). This data privacy failure is indicative of a failed data privacy management program where personal data is protected from end to end in a formal data lifecycle. This paper will discuss elements that are critical to the success of a data privacy management program with a specific emphasis on data mapping.

Data Mapping Requirement

The California Consumer Privacy Act (CCPA) and General Data Protection Regulation (GDPR) mandate for detailed data inventories are now reflected in many other privacy regulations worldwide and continue to pose a significant challenge for organizations of all sizes (Gabrielian, 2023). It is important to note that the GDPR does not explicitly use the term "data mapping" but instead requires multiple components that constitute the activity of data mapping (Sullivan, 2022). Specific requirements that are supported by a healthy data map include Article 30 creating and maintaining a records of processing activity (ROPA), Article 33 Breach Management, Article 35 Conducting data protection impact assessments (DPIA), and fulfilling privacy requests (Sullivan, 2022). Similarly, under CCPA there is not a explicit data mapping requirement but a collection of requirements that are fulfilled by proper data mapping. Section 1798.130 of CCPA specifically requires organizations to satisfy a consumer's request to disclose all personal information collected, sold or shared in the previous 12 months (California Legislature, 2018).

Privacy Data Mappings BY: Professor Jordan A. VanHoy

Further, Section 1798.120 of CCPA forbids the organization from selling personal information of consumers without providing a notice to consumers and providing consumers with a “right to opt out” of the sale (California Legislature, 2018). Knowing what data constitutes personal information for their consumers are answered by performing data mapping.

Taking a further look into these articles we see that article 30 stipulates that the record shall contain the following information: the contact information of the data controller and joint data controller as appropriate, purpose(s) of processing the data, descriptions of the categories of data subjects and categories of personal information, categories of recipients to the personal information, information transfers, data retention limits on the categories of information, and descriptions of security measures placed on the information (GDPR, 2018, Art. 30). This article alone demonstrates the necessity of proper data mapping. Data mapping is about identifying the data you have, how the data flows across the organization, how the data is used, why it used, how long it is kept for, and who the intended recipient of that data is to be. Therefore, by conducting a comprehensive data mapping organizations can fully comply with Article 30 of the GDPR.

Article 33 deals with breach management under the GDPR. Within the regulation article 33 mandates that in the event of a data breach the data controller must notify the supervisory authority within 72 hours (GDPR, 2018, Art. 33). Supervisory authority in the context of GDPR refers to a independent public authority responsible for monitoring compliance with GDPR and is designated within each member state of the European Union (GDPR, 2018, Art. 51). Where a proper data mapping can assist the data controller is by satisfying the requirements of describing the categories of data breached, the number of data subjects and records involved, and the likely consequences involved with inadvertent exposure of the records.

Finally, Article 35 of GDPR covers the conduct of a data protection impact assessment. This is to occur when a high risk to the rights and freedoms of natural persons may occur due to data processing (GDPR, 2018, Art. 35). The DPIA is a systematic and extensive evaluation of personal aspects related to individuals when their data is subject to automated processing, profiling, and when legal decisions are made that impact the individual (GDPR, 2018, Art. 35). Understanding the categories of data and how much data falls under those categories is fundamental in being able to complete a DPIA. While not explicitly stated that data mapping is required, it is evident that data mapping fulfills so many unique requirements for GDPR.

CCPA is a groundbreaking privacy reform in the United States that was directly influenced by the GDPR landscape. CCPA is a standalone regulation that shares similarities with GDPR but also explicitly requires actions to be taken that GDPR does not. Of the notable CCPA requirements satisfied by data mapping is Section 1798.130. Users must have two or more designated methods for requesting information on their records and for such issues that may arise as deletion or correction (California Legislature, 2018). This also applies to records applicable to the user that have been sold or shared within the last 12 months. By having a mature mapping process in place, it reduces the burden of responding to such requests. Section 1798.120 allows users to “opt-out” of having their data sold or shared with third parties (California Legislature, 2018). This is a stark contrast to the GDPR as the GDPR defaults to a more secure “opt-in” process. The significance of data mapping is on display when organizations write their privacy policies and notices regarding the sale or sharing of data. Without understanding the type and quantity of data transferred it is nearly impossible to develop an accurate notice or policy.

Privacy Data Mappings BY: Professor Jordan A. VanHoy

Data Mapping Framework

A framework is a basic conceptual structure of interrelated ideas. Applied to privacy we see that privacy program management is the structured approach of combining several projects into a framework and life cycle to protect personal information and the rights of individuals (Densmore, 2022). Two frameworks used to conduct data mapping are Privacy by Design and the Privacy Governance Lifecycle. Each framework can be viewed as a wholistic approach to privacy rather than a specific approach to data mapping itself. While data mapping is a very important component of a healthy privacy program, it is certainly not the only function.

The International Association of Privacy Professionals (IAPP) has designated an approach to privacy program management by utilizing the Assess, Protect, Sustain, Respond model (Densmore, 2022). Within the assess function of the framework organizations are charged with obtaining or developing the steps, checklists, and processes necessary to assess gaps in a privacy program as compared to best practices, regulations, and company policies (Densmore, 2022). The IAPP framework allows for flexibility similar to the National Institute of Standards and Technology Cybersecurity Framework where other privacy frameworks can be used to satisfy requirements for this first phase. Additional frameworks that may be used to develop the assess portion of this phase is the AICPA/CICA Privacy Maturity Model, Generally Accepted Privacy Principles (GAPP), and Privacy by Design (PbD). The deliverable in this phase is to perform a gap analysis, understand the who, what, when, where, why, and how of the data collected.

During the protect phase, the organization has identified the data and data categories and is looking to secure the information. An eloquent dance between information security and privacy are shared as they are mutually supporting but conceptually different. Information can be secure and not private just as information can be private and not secure. Hence privacy, cybersecurity, and information assurance must work in tandem to achieve a wholistic protection that is further complicated across jurisdiction, compliance, and law (Densmore, 2022). Each security related tool employed across an organization helps the privacy program meet its requirements.

The third phase in the IAPP framework is sustain. Sustain takes an organization through the monitoring, auditing, and communication aspects of the management framework (Densmore, 2022). This includes incorporating automation to the greatest extent possible in order to alleviate pressure placed on manual tasks performed by privacy staff. This is particularly important as it applies to data mapping in dynamic environments. As additional data sources are brought online the data should be classified, have data flows documented, and risk analysis performed to ensure appropriate controls are applied by both privacy and security teams alike.

The final phase of the IAPP framework is to respond. This phase includes the respond principles of information requests, legal compliance, incident-response planning, and incident handling (Densmore, 2022). Every organization needs to be prepared to respond to its customers, partners, vendors, employees, regulators, shareholders, or other legal entities (Densmore, 2022). These requests may come in many forms to include requests on the types of data collected on an individual, data corrections, data deletions, and data sharing requests. Having a fully functioning privacy program will ensure transparency and visibility into how the organization responds to these types of requests. Some of which are scrutinized by regulations such as GDPR and CCPA.

Privacy Data Mappings BY: Professor Jordan A. VanHoy

Privacy by Design is a approach developed by Dr. Ann Cavoukian the Information & Privacy Commissioner for Ontario, Canada. Predicated on seven principles, the PbD framework aims to provide a standard that is higher than what the Fair Information Practices (FIP) achieved. The seven principles are:

- Proactive not Reactive; Preventative not Remedial
- Privacy as the Default Setting
- Privacy Embedded into Design
- Full Functionality – Positive-Sum, not Zero-Sum
- End-to-End Security – Full Lifecycle Protection
- Visibility and Transparency – Keep it Open
- Respect for User Privacy – Keep it User-Centric (Cavoukian, 2014).

The first principle is something we all seek to achieve in cybersecurity. Instead of waiting for risks to materialize we aim to prevent them before they occur. The second principle is predicated on ensuring data is automatically protected and does not require action on behalf of the user to initiate protection of their data. Privacy embedded into design and architecture of IT systems and business practices means that privacy is not thought of as a bolt on solution but rather initiated at the onset. Privacy is integral to the system, without diminishing functionality (Cavoukian, 2010). Full functionality touts that unnecessary tradeoffs do not need to occur in order to provide for privacy. The impacts of privacy and related controls should be demonstrably minimized. Further, security is a critical component to providing privacy. This ensures that all data will be securely retained, and then securely destroyed at the end of the process and in a timely fashion (Cavoukian, 2014). The sixth principle reinforces confidence in the stated protections to the appropriate stakeholders by making components of the program visible and transparent. Finally, architects and operators must keep privacy options user centric. Empowering data subjects to play an active role in the management of their data is an effective and necessary function to prevent abuse and mismanagement.

Data Mapping Techniques

There exist multiple methods for conducting data mapping whether it be automated or manual. Data sources, data types, and data classifications are among many features that require a flexible method of conducting data mapping. Data may be manipulated in endless possibilities and choosing the appropriate technique for mapping will ensure a positive and sustainable privacy management program.

Direct Mapping

Direct mapping can be used manually or via automated tools. This is however the most primitive manner of conducting data mapping and generally done when the privacy program is in a immature state. When conducting the direct mapping technique, data owners map data fields directly from source to target without any transformations or modifications (Noss, 2023). This method is analogous to providing equal protection of assets within the organization instead of focusing on a data centric protection model.

Concatenation

Concatenation can be used manually or via automated tools. This technique involves combining multiple data fields from the source into a single target field (Noss, 2023). This can be especially useful when combining fields that contain related data into a single field. For example, a user may want to take the fields of last name and first name to combine them into a full name field.

Privacy Data Mappings BY: Professor Jordan A. VanHoy

Lookup Tables

Lookup tables can be used manually or via automated tools. This is used when replacing or mapping source data values to corresponding values in the target system using a lookup table (Noss, 2023). Two popular Microsoft Excel functions called HLookup and VLookup are used in general productivity and provide a semi-automated method of using lookup tables. By leveraging a lookup table one is able to map data values by looking in a single row or column and the need to find a value from the same position in a second row or column (Microsoft, 2023).

Data Transformation

Data transformation can be used manually or via automated tools. Transformation involves Manipulating or converting data during the mapping process, such as cleansing, aggregation, or calculations (Noss, 2023). This is a very common technique used to normalize data and present the information in a manner that enhances the user interface. Common data transformation examples include converting a Microsoft Word document to Adobe PDF and taking raw machine data from a Security Information and Event Management (SIEM) tool to build dashboards of visual context.

Conditional Mapping

Conditional mapping can be used manually or via automated tools. In more complex environments it may be beneficial to conduct mapping between source and target fields based on specific conditions or rules (Noss, 2023). When data classifications are introduced into a environment, conditional mapping may be one of the only techniques viable to ensure both privacy and security. Conditional mapping may also be analogous to the Bell-Lapadula model in security where subjects and users are organized according to differing layers of security.

Scheme Mapping

Scheme mapping can be used manually or via automated tools. This technique has the data owner mapping the overall structure or schema of the source data to the target data structure (Noss, 2023). This is used when the source scheme and target scheme have structural differences that must be joined for a final product. This is a common occurrence given that each schema includes unique feature types, permitted geometries, user-defined attributes and other rules that define or restrict its content.

Field-Level Mapping

Field level mapping can be used manually or via automated tools. When source and target schemas align, field level mapping can be done by mapping individual data fields from the source to their corresponding fields in the target system (Noss, 2023). This technique allows data owners to eliminate end users placing data in multiple places and is focused on ensuring data quality remains high.

Hierarchical Mapping

Hierarchical mapping can be used manually or via automated tools. When data must be categorized from a top down manner this technique is especially beneficial. The mapping process is complete by mapping parent-child relationships and maintaining hierarchical structures in nested data formats (Noss, 2023). This process begins in a manner similar to a funnel by moving from generalized types of data to specific types of data.

Privacy Data Mappings BY: Professor Jordan A. VanHoy

Semi-Automated Mapping

Semi-automated mapping can be used manually or via automated tools. One of the more efficient methods listed this seeks to combine manual mapping with automated suggestions or recommendations from machine learning algorithms (Noss, 2023). In computing there is an element of human judgement needed to ensure a proper outcome is produced. Using a semi-automated method can be beneficial when starting a privacy program as the user will have a greater influence into how the tool will map data elements.

Template Based Mapping

Template based mapping can be used manually or via automated tools. A common methodology in security and privacy is to leverage best practice templates. Template based mapping accomplishes this by using pre-defined templates or mappings for commonly encountered mapping scenarios (Noss, 2023). For programs that need to comply with specific regulations this can be a significant time enhancement and produce consistent results.

Metadata Driven Mapping

Metadata driven mapping can be used manually or via automated tools. In a rather technical and detailed manner the metadata driven mapping utilizes metadata about the source and target data to guide the mapping process (Noss, 2023). Metadata is information about the information being used. An example of this would be the metadata in a email header that describes the sender, recipient, date, subject, and attachments used.

Automation of Data Mapping

Manually conducting data mapping is time intensive, does not provide a proper return on investment, and may lead to data management problems in the future. In an analogous manner, cybersecurity has touted the transition from static audits to a concept known as information system continuous monitoring (ISCM). To assist in this the National Institute of Standards and Technology has developed the NIST Interagency Report (IR) 8011 series with volumes one through four. NISTIR 8011 is Automation Support for Security Control Assessments and seeks to provide guidance for organizations to effectively use automation to collect evidence used in control assessments. Similarly, tools have paved the way for automation to enter the privacy realm.

While not a data cataloging tool, Microsoft launched a new open source tool for mapping multiple privacy laws and their inter-related standards in a easy to use interface. The tool maps ISO/IEC 27701 to the EU General Data Protection Regulation, California Consumer Privacy Act, Brazil's General Data Protection Law, Australia's Privacy Act, Canada's Personal Information Protection and Electronic Documents Act, Singapore's Personal Data Protection Act, Hong Kong's Personal Data Ordinance, South Korean's Personal info Protection Act, and Turkey's Data Protection Law (Fennessy, 2022). Until this tool had been created, organizations were left to map requirements on their own. With new privacy laws and regulations popping up almost daily this became a nightmare for organizations to track. Laying the groundwork for data mapping begins in the assess phase where understanding of applicable laws and regulations takes place.

Taking a step further, tools have also been developed to conduct automated data mapping for organizations. A look at Transcend Data Mapper shows that the tool can discover data silos, classify personal data, and auto-generate reports – all in an easy-to-use, collaborative platform (Transcend, 2023). Reducing as much human element as possible, Transcend uses a scanning technology to integrate into the environment and company website to auto populate results into the data inventory. A feature that helps set this tool apart from competition is its ability to use an engine to “smartly” classify data and categorize personal data points. In another separation from competition, Transcend touts core principles such as “secure-by-design” architecture which provides privacy to the extent that Transcend cannot see the organizations data nor does it have direct access to the organizational systems.

Privacy Data Mappings BY: Professor Jordan A. VanHoy

Clarip is another heavyweight in the privacy data mapping field and touts that it is the most powerful platform for managing data privacy of the 24/7 connected customer (Clarip, 2023). A hallmark of this product is Clarip's ability to harness the power of Artificial Intelligence (AI) and Machine Learning (ML) to make the data mapping process as seamless as possible on the end user. By using the build in analytics of Clarip, organizations can benefit from more meaningful insights while making compliance painless. An alternate aspect of efficiency provided by Clarip is the ability to obtain consent and preference management through a unified and simple interface to bring enhanced customer engagement. As with any tool leveraging AI/ML the tool is constantly scanning the environment to detect changes or ingestion of new data and simultaneously having that data mapped appropriately providing ease of mind.

Onetrust provides fully automated data mapping features that specifically addresses an organizations requirement of maintaining ROPA. Similar to Clarip, Onetrust seeks to provide constant coverage of what is coming into the environment. A significant benefit of Onetrust is the always available and up to date ROPA that can be pulled for compliance checks. Where Onetrust does not shine is the data ingestion process where data owners are presented with a series of questionnaires that facilitate the ingestion process (Onetrust, 2023). Clarip does not require the data owner to fill out questionnaires for data ingestion. However, Onetrust has a built in Privacy Impact Assessment and DPIA function that is designed to automate the entire assessment. The process of creating, distributing, and analyzing PIAs and DPIAs requires automation to effectively and efficiently achieve privacy by design (Onetrust, 2023).

Conclusion

Despite recent regulation aimed at improving user privacy, organizations continue to make fundamental privacy errors as seen in the case against Doorstep Dispensaree. This organization dumped approximately 500,000 personal health information records in a public location. The lack of care for these personal records demonstrates the immaturity of the privacy management program and can benefit from implementing techniques such as data mapping. By leveraging data mapping, organizations can categorize and classify information to assist in end-to-end protection of information. This begins with understanding where the data sources are, classifying the data, understanding data flows across the organization, analyzing the risks associated with the data, and documenting this in a manner that is easy to sustain. Following recognized frameworks like the IAPP assess, secure, protect, respond can help provide structure to a growing privacy program. Alternate philosophies like privacy by design seek to build privacy in every function of an organization without compromising on functionality. CCPA and GDPR are two regulations among many that drive organizations to complete data mapping for their organizations. While not explicitly required under either regulation, data mapping provides a foundation from which many of the compliance requirements may be answered with ease. Additionally, data mapping is a cornerstone for a effective privacy management program. Understanding where data is coming from, the sensitivity of that data, and where the data flows across the organization creates an effective blueprint for how to secure information along the way. Without an effective baseline there is no manner of detecting anomalies or infractions. The most efficient manner of conducting data mapping is to automate as much as possible to eliminate human error and reduce manual effort. Multiple tools exist to automate data mapping to include Clair, Transcend, and Onetrust. While each program has their advantages, Onetrust provides a unique advantage by automating PIA and DPIA assessments for organizations. This is significant as DPIA's are a requirement under GDPR among many other requirements that data mapping can answer therefore effectively handling two significant components of GDPR compliance.

Privacy Data Mappings BY: Professor Jordan A. VanHoy

References

- Boger, S. (2020). GDPR Case Study: Doorstep Dispensaree Ltd. Brown University. <https://cs.brown.edu/courses/csci2390/2020/assign/gdpr/sboger-doorstep-dispensaree.pdf>
- California Legislature. (2018). California Law Code Section https://leginfo.ca.gov/faces/codes_displaySection.xhtml?lawCode=CIV§ionNum=1798.130.
- Cavoukian, A., (2014). Keynote - Ann Cavoukian - The Need for Big Privacy in a World of Surveillance and Big Data. YouTube. Retrieved August 29, 2023, from <https://www.youtube.com/watch?v=MK9eZB0fsiM>.
- Cavoukian, A. (2010). Privacy by Design The 7 Foundational Principles Implementation and Mapping of Fair Information Practices. International Association of Privacy Professionals. https://iapp.org/media/pdf/resource_center/pbd_implement_7found_principles.pdf
- Clarip, Inc. (2023). Clarip Privacy Governance - Business. <https://www.clarip.com/business>
- DataGuidance. (2022). UK: ICO fines Doorstep Dispensaree £275,000 for failure to ensure. <https://www.dataguidance.com/news/uk-ico-fines-doorstep-dispensaree-%C2%A3275000-failure>
- Densmore, R. (2022). Privacy Program Management: Tools for Managing Privacy Within Your Organization (3rd ed.) [E-pub]. International Association of Privacy Professionals (IAPP).
- ESRI. (2023). Understanding schema. https://webhelp.esri.com/arcgisdesktop/9.3/datainterop/mergedProjects/FME_Workbench/understanding_schema.htm#:~:text=want%29.%20Schema%20mapping%20is%20the%20process%20of%20connecting,attributes%20are%20sent%20to%20the%20right%20destination%20attributes.
- Fennessy, C. (2022). Microsoft launches open-source privacy mapping tool. International Association of Privacy Professionals. <https://iapp.org/news/a/microsoft-launches-open-source-privacy-mapping-tool/>
- Gabrielian, Y. (2023). Data Inventory Fundamentals - Initial Steps to take [Video]. Kroll. <https://www.kroll.com/en/services/cyber-risk/assessments-testing/data-mapping-gdpr-ccpa>
- Microsoft. (2023). LOOKUP function. <https://support.microsoft.com/en-us/office/lookup-function-446d94af-663b-451d-8251-369d5e3864cb>
- Noss, S. (2023). Data mapping process steps & workflow for GDPR compliance. DataGrail. <https://www.datagrail.io/blog/data-privacy/data-mapping-101-purpose-process-tools/>
- OneTrust. (2023). Data Mapping Automation | Products. <https://www.onetrust.com/products/data-mapping-automation/>
- Sullivan, M. (2022). The complete Guide to GDPR data Mapping. transcend.io. <https://transcend.io/blog/gdpr-data-mapping/>
- Transcend. (2023). Automated Data Mapping | Transcend. <https://transcend.io/data-mapping/>

Welcome to the SEPTEMBER 2023 issue of THE PACKET! As we usher in the fall semester, the University of Arizona Cyber Operations program is primed to bring you the latest from the cybersecurity frontier. I'm your ever-present guide, Professor Michael Galde. I am delighted to welcome our returning scholars and fresh faces to an academic year brimming with insights, challenges, and milestones.

While many basked in the summer sun or trekked scenic trails, the digital realm was ablaze with activity, thanks to relentless cyber adversaries. Our news section delves into some of these incidents, including the Duolingo Debacle and the machinations of Flax Typhoon. I've also shed light on various cyberattacks, infusing them with geopolitical context for a richer understanding. Digging deeper, my article on the Windows Filtering Platform (WFP) offers a meticulous exploration of its vulnerabilities and potential for exploitation, revealing advanced attack techniques and potential security workarounds.

In our ceaseless quest to navigate the intricate landscape of cybersecurity, we continually scout for tools and pioneering projects. This month, we're spotlighting two intriguing endeavors: SearXNG and PrivateGPT. Embodying the essence of innovation and adaptability in cybersecurity, these projects are a testament to the ever-evolving nature of our field. For all the budding professionals and seasoned experts out there, delving into such initiatives augments your expertise and positions you at the vanguard of tech progression. Immerse, tinker, and cultivate an insatiable thirst for knowledge!

As we collectively set forth on this academic odyssey, I cannot stress enough the vitality of staying attuned to the ever-shifting sands of cybersecurity. Engage, debate, and dissect. The digital realm beckons the astute minds among you to champion its safeguarding. So, here's raising a toast to a semester filled with revelations, hurdles, and triumphs. Stay inquisitive, remain vigilant, and let's sculpt an unforgettable year together!

CONTACT US

CIIO@EMAIL.ARIZONA.EDU

1140 N. Colombo Ave. | Sierra Vista, AZ 85635

Phone: 520-458-8278 ext 2155

<https://cyber-operations.azcast.arizona.edu/>

